NORMSAGE: Multi-Lingual Multi-Cultural Norm Discovery from Conversations On-the-Fly

Yi R. Fung[1], **Tuhin Chakraborty**[2], **Guo Hao**[3]
Owen Rambow[4], **Smaranda Muresan**[2], **Heng Ji**[1]
[1]University of Illinois Urbana Champaign, [2]Columbia University
[3]Tsinghua University, [4]Stony Brook University
{yifung2,hengji}@illinois.edu, smara@cs.columbia.edu

Abstract

Norm discovery is important for understanding and reasoning about the acceptable behaviors and potential violations in human communication and interactions. We introduce NORMSAGE[1], a framework for addressing the novel task of conversation-grounded multi-lingual, multi-cultural norm discovery, based on language model prompting and self-verification. NORMSAGE leverages the expressiveness and implicit knowledge of the pretrained GPT-3 language model backbone (Brown et al., 2020a), to elicit knowledge about norms through directed questions representing the norm discovery task and conversation context. It further addresses the risk of language model hallucination with a self-verification mechanism ensuring that the norms discovered are correct and are substantially grounded to their source conversations. Evaluation results show that our approach discovers significantly more relevant and insightful norms for conversations on-the-fly compared to baselines ($\geq$10$^+$% in Likert scale rating). The norms discovered from Chinese conversation are also comparable to the norms discovered from English conversation in terms of insightfulness and correctness ($\leq$ 3% difference). In addition, the culture-specific norms are promising quality, allowing for 80% accuracy in culture pair human identification. Finally, our grounding process in norm discovery self-verification can be extended for instantiating the adherence and violation of *any* norm for a given conversation on-the-fly, with explainability and transparency. NORMSAGE achieves an AUC of 95.4% in grounding, with natural language explanation matching human-written quality.

1 Introduction

Norms are rules that embody the shared standards of behaviors amongst cultural groups and societies (Abrams et al., 2022). These may include *social conventions* (*e.g.,* it's good to shake hand with your opponent even if you lost); *behavior guidances* (*e.g.,* it's wrong to hurt a pet); or *general concepts* (*e.g.,* it's nice to be smart) (Forbes et al., 2020; Ziems et al., 2022). Along this direction, the SOCIAL-CHEM-101 (Forbes et al., 2020) and MORAL INTEGRITY CORPUS (Ziems et al., 2022) present two manually annotated, rule-of-thumb[2] catalogues.

However, current norm discovery approaches come with two major shortcomings. First, the approaches are primarily based on manually constructing a static norm library from curated English data, such as Reddit post headers and Dear Abby column titles (Forbes et al., 2020; Ziems et al., 2022; Gu et al., 2022). This process not only is time-consuming and expensive, but also limits the portability of the discovered norms in human interaction understanding across data domains and sociocultural[3] groups. For example, while a Reddit post header may mention general long work hours in developing countries, deeper norms, such as the acceptability of night shifts for jobs outsourced to `India`, are likely found from details in conversations on-the-fly instead. Secondly, while there exist some preliminary explorations on generating norms from titles, based on the static norm library annotated (Forbes et al., 2020; Ziems et al., 2022), they fail to quality control over correctness and insightfulness. Such norm discoveries suffer from over-dependence on title-like text as source of data, and cannot handle in-situ conver-

[1]We will publicly release our code, data, and Github repository upon publication.

[2]"Rule-of-thumb" and "norm" are synonymous references.

[3]Previous work explored norms across "cultures" in terms of moral foundations (Forbes et al., 2020; Ziems et al., 2022), but overlooked the shared socioethnic beliefs and behaviors.

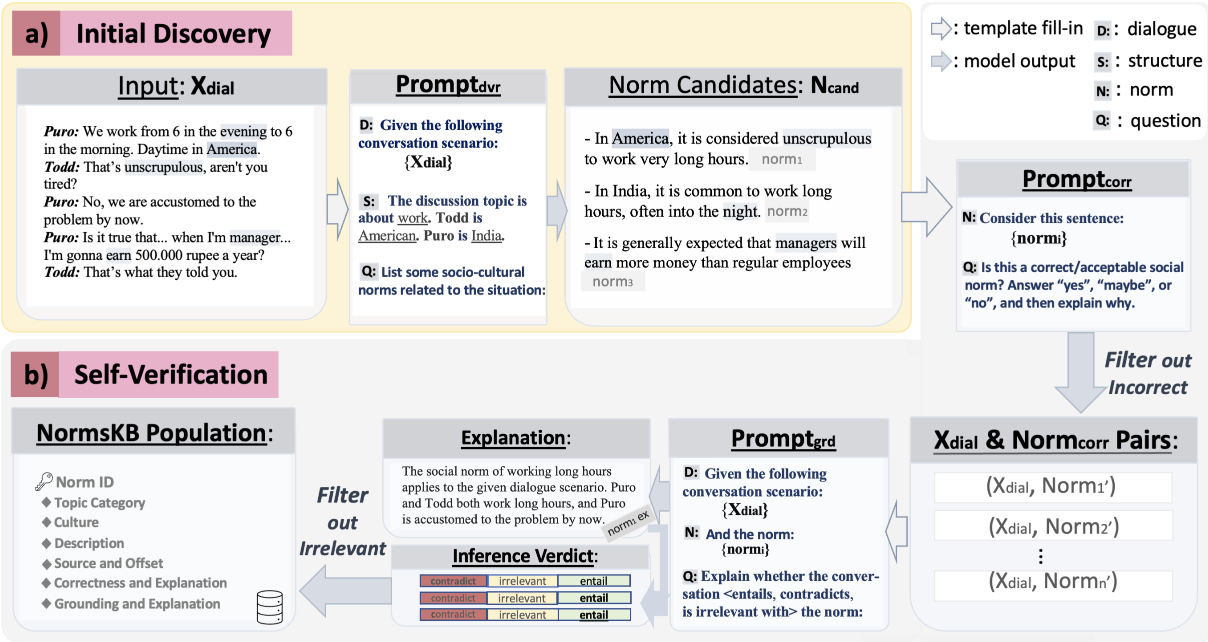

Figure 1: The key idea of NORMSAGE is prompting & verification for conversation-grounded norm discovery.

sations well. Their knowledge about norms also scales poorly for unseen topics, specific cultures, and data beyond English language sources.

In this paper, we explore a new paradigm for direct norm discovery from conversations on-the-fly, and address out-of-domain concerns by sidestepping the need for human annotation. We propose eliciting knowledge (Petroni et al., 2019) about norms from large pretrained language model through prompting (Liu et al., 2021) and self-verification. In particular, the GPT-3 (Brown et al., 2020b) model has been pretrained on 45TB of text across various datasets and the web, with strong zero-shot capabilities across a variety of tasks. We take advantage of the implicit knowledge in GPT-3, and operationalize prompting the model for norm discovery over dialogue situations. We show that surprisingly relevant norm descriptors can be derived simply by feeding a dialogue exchange, in any of the major languages (*e.g.*, English, Chinese, etc.), through GPT-3 along with a direct question inquiring for norms related to the situation. These descriptors may lack form (*i.e.*, judgement on norm acceptability or occurrence) and taxonomy (*i.e.*, categorization on the topic, culture, etc.). But we demonstrate that structural priming, with the inclusion of *frame guidelines* (*e.g.*, description of the expected format of norms in the prompt) or *extracted indicators* (*e.g.*, speaker ethnicity, gender, age, profession, etc.) in the natural language prompt template, effectively enforce well-formedness and taxonomy in the norm discovery.

To safeguard against discovering norms corrupted from biased conversations or language model hallucinations, we further propose a self-verification formulation. Specifically, we develop a duo filtering mechanism to quality-check each discovered norm for correctness and insightfulness. We extend the prompting operation with the same language model, GPT-3, from discovering norms, to making sure that the norms hold in society and can be substantially grounded back to (*i.e.*, entailed or contradicted by) the source dialogue.

Our contributions can be summarized as the follows. I) We define the novel task of conversation-based, multi-lingual multi-cultural norm discovery. II) We propose NORMSAGE, a zero-shot language model prompting and self-verification framework, which discovers norms rated $\geq$ 10% more insightful than baselines. The culture-specific norm discoveries are also appropriate and expressive, allowing for 80% accuracy in cross-culture binary identification assessment. III) In addition, our self-verification mechanism that grounds norm candidates discovered back to dialogue sources extends for instantiating whether a norm is adhered or violated by *any* given dialogue. We achieve probability scorings, with an AUC of 95.4%, and natural language explanations comparable to human-written ones. IV) Finally, we present a large corpus of multi-lingual multi-cultural source conversations ($\geq$1.5M tokens), along with the **NORMKB** discovered, as an asset to the community.

2 Details of the Source Data

In collecting source data for norm discovery, we have the following selection criteria. First, the data should involve in-situ conversations to best mimic or reflect real-world communications. This is because we want to enable direct norm discovery from human-human interactions on-the-fly, instead of a limited set of curated media summaries. Secondly, the data should ideally span diverse topics and societal or cultural groups. Yet, it is generally difficult to obtain large-scale, real-world data for norm discovery due to privacy concerns, as well as sparsity of interesting human interaction occurrences. Thus, we expand on the predominantly single-cultured TVQA dataset (Lei et al., 2018), and collect a set of TV, movies, and documentaries covering different cultures, as detailed in Table 1. Finally, we include several multi-lingual conversations from real-world negotiations, chats, and documentaries to explore norm discovery adaptability in diverse data settings.

	Source of Data	# Tok	# Ln
Single Culture	Big Bang Theory	29,682	3,468
	Friends	26,197	2,849
	How I Met Your Mother	29,423	3785
	Grey's Anatomy	23,341	3,117
	Castle	38,880	4,142
Cross-Culture	Fresh off the Boat	26,056	4,129
	Never Have I Ever	39847	5,637
	Blackish	33,993	5,103
	Citizen Khan	22,985	3,198
	Outsourced	1464	123
Multi-Ling.	American Factory Documentary	10,840	1,138
	Real-World Negotiations	19,487	1,758
	LDC CCU TA1 Chinese Dev.	1.2M	102k
	Total	1.5M	140k

Table 1: Sources of raw data and their statistics, including the number of tokens (tok) and lines (ln).

3 Methodology

3.1 Task Formulation

Our overarching goal is to derive a knowledge base of norms (**NORMSKB**) that can be dynamically updated based on conversations on-the-fly to help reason about acceptable behaviors and common expectations across cultural and language groups in the world. With this in mind, we define the **conversation-based, multi-lingual multi-cultural norm discovery** problem as follows. Given a conversation scenario ($\mathbf{X}_{dial}$) in one of the pre-defined target languages (*e.g.*, English, Chinese, etc.), we aim to utilize an automatic norm discovery framework (*e.g.*, **NORMSAGE**) to derive a list of norms $\mathbf{N} = [\mathbf{n}_1 ... \mathbf{n}_m]$, which can be used to populate the **NORMSKB** library. In practice, the conversation scenario ($\mathbf{X}_{dial}$) can be preprocessed into chunks ($\mathbf{X}_{dial_{1..N}}$), each consisting of a certain number of dialogue exchanges, to account for language model maximum token length constraints and encourage information processing granularity in the norm discovery. We set this number arbitrarily to $|\mathbf{X}_{dial_i}|_{\#lines} = \mathbf{5}$, and derive a set of candidate norms ($\mathbf{N_{cand}}$) from each dialogue chunk ($\mathbf{X}_{dial_i}$).

As handling hallucination and achieving transparency are important for norm discovery, we introduce the **supplementary verification task** of checking norm correctness and relevance. We aim to filter out incorrect norms, by deriving a correctness verdict $\mathbf{C}_v \in \{\mathbf{1} : yes, \mathbf{0} : maybe, \textbf{-1} : no\}$, along with a confidence probability ($\mathbf{C}_p$) and natural language explanation ($\mathbf{C}_{expl}$). Additionally, we aim to filter out non-insightful norms, by deriving a grounding inference $\mathbf{G}_v \in \{\mathbf{1} : entail, \mathbf{0} : irrelevant, \textbf{-1} : contradict\}$, along with a confidence probability ($\mathbf{G}_p$) and natural language explanation ($\mathbf{G}_{expl}$). Candidate norms with low correctness or grounding relevance score will be filtered out. The remaining norms ($\mathbf{N^{fil}_{cand}}$), along with their corresponding dialogue examples, correctness information $\mathbf{C} = (\mathbf{C}_v, \mathbf{C}_p, \mathbf{C}_{expl})$, and grounding information $\mathbf{G} = (\mathbf{G}_v, \mathbf{G}_p, \mathbf{G}_{expl})$, will be added to the **NORMSKB**.

This conversation-grounded, multi-lingual multi-cultural norm discovery task is novel in several key aspects. It is the first task to define automatically discovering norms from dialogue data, which best reflects in-situ human communication. In addition, it is the first task to define discovering norms from multi-lingual domain, and discovering norms with culture-specificity on the shared practices and beliefs within socioethnic groups. Grounding the discovered norm with dialogue examples, confidence score, and natural language explanations is also new, allowing for norm discovery to be explainable and self-supervised. From a larger picture, this verification process benefits downstream application and human users as well because it naturally extends to determine whether a conversation scenario adheres to or violates any norm.

3.2 The NormSage Framework

We propose **NormSage**, a language model prompting and self-verification framework for discovering norms from conversations on-the-fly.

Core Approach: Pretrained language models (PLMs) store implicit knowledge about the world learnt from large-scale text collected around the internet (Petroni et al., 2019). We frame conversation-based norm discovery as a series of natural language prompts, each with a directed question for the pretrained GPT-3 Davinci[4] language model to reason with its internal knowledge and generate an answer response. To discover an *initial* set of candidate norms from conversation data, we introduce the $\textbf{PROMPT}_{\textbf{DVR}}(.)$ operator, which concatenates:

D – a template header describing the nature of the context data, followed by a fill-in slot $\{\mathbf{X}_{\mathbf{dial}_i}\}$ for the actual dialogues;

Q – a directed question describing the norm discovery task

as input for the PLM to generate response. Because multiple discrete questions (**Q**) may apply for describing the task to prompt norms, producing complementary results, we re-run probing through the PLM for each variation of prompt template shown in Table 2 to expand on the list of norm candidate discoveries ($\mathbf{N}_{\mathbf{cand}}$).

Structure Enhancement: A shortcoming observed in standard prompting is that the norms discovered may lack well-formedness and taxonomy for categorizing information specific to different cultures and topics. To encourage greater level of detail and structure in $\textbf{PROMPT}_{\textbf{DVR}}(.)$ outputs, we add to the prompt input:

Question (Q) in Prompting	Example Norm Discovery Output
*"What are some **socio-cultural** norms related to the situation?"*	*"It is normal to have a set **schedule for work**"*
*"What are some **moral** norms related to the situation?"*	*"It is **wrong to lie** to workers about their compensation"*
*"List some **social** norms and advice related to the situation:"*	*"It is considered **polite** to **inquire about** someone's **well-being** before diving into **conversation**."*

Table 2: For the same dialogue input from Fig. 1, different question for prompting (socio-cultural/moral/social) lead to different categories of norm discovery.[6]

S – a building block in the text template consisting of either frames defining the expected structure of norms, or structured indicators such as:

- The *discussion topic*, which can be extracted by prompting on the dialogue (*e.g.*, *"what is the overall discussion topic in the conversation scenario $\{\mathbf{X}_{\mathbf{dial}_i}\}$?"*)
- Speaker *ethnicity*, *gender*, *age group*, and *profession*, depending on information availability. These indicators can be extracted through similar prompting on the Wikipedia background summary of the movies, shows, and high-profile meetings for example.

We see in Fig 2 how structured indicators encourage a culture-specific and topic-specific taxonomy in the norm discovery process, which helps information categorization in **NormsKB** population.

Self-Verification with Correctness Checking & Explainable Grounding:

For each of the norms discovered, we add a $\textbf{PROMPT}_{\textbf{corr}}(.)$ operator to check the correctness of norms. This prompting operator follows the natural language template of: *"Consider this sentence: $\{\mathbf{n}_i\}$. Is this a correct/acceptable social norm? Answer '**yes**', '**maybe**', or '**no**', and then explain why."*. The output from $\textbf{PROMPT}_{\textbf{corr}}(.)$ consists of both a correctness verdict, $\mathbf{C_v}$ and a

[4] There is a convenient API available at https://openai.com/api/

[6] Note: the term *"socio"*, in "**socio**-cultural" refers to general society, while the term *"social"*, in "**social** conventions", refers to socializing etiquettes (*e.g.* talking and intermingling).

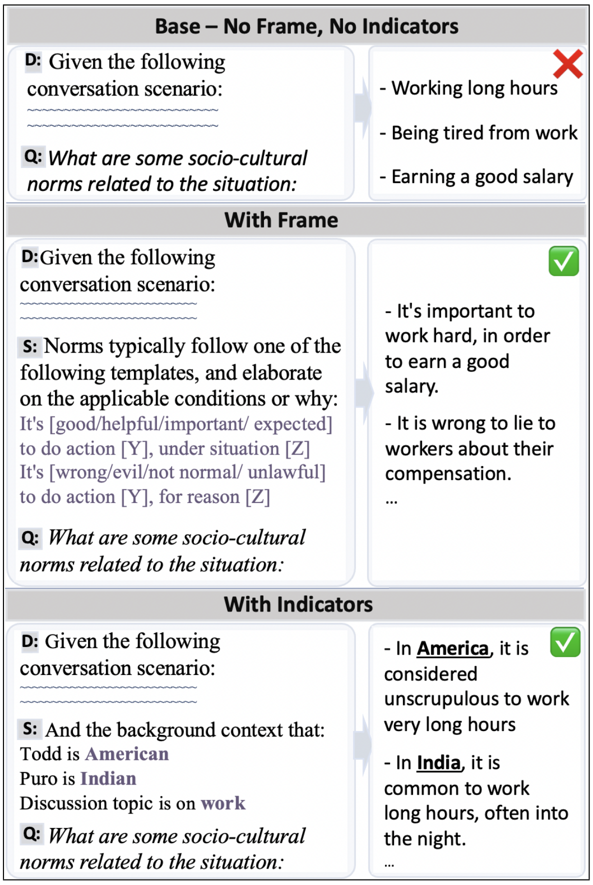

Figure 2: A comparison of the variants of NORMSAGE, in *base* form without structure enhancement (top), *with frame* (middle), and *with indicators* (bottom) to guide the norm discovery.

subsequent explanation, $\mathbf{C_{expl}}$, in a single natural language response generation. As an example,

> $\hat{\mathbf{C}}_\mathbf{v}$: *"Yes"*
> $\hat{\mathbf{C}}_\mathbf{expl}$: *"This is a correct/acceptable social norm because it is considered unscrupulous to work very long hours in America."*

We further derive a confidence score for the correctness verdict by normalizing the probability of token generation for $\mathbf{C_v}$ ='yes' over the probability of the alternative maximum likelihood prediction:

$$\hat{\mathbf{C}}_\mathbf{p} = \frac{\mathbf{P}(\mathbf{C_v} = \text{`}yes\text{'})}{\mathbf{P}(\mathbf{C_v} = \text{`}yes\text{'}) + \mathbf{P}_{\max}(\mathbf{C_v} \neq \text{`}yes\text{'})} \quad (1)$$

Norm candidates with correctness probability below a tunable threshold, of $\theta = 0.7$, are filtered out, with the remaining norm candidates following:

$$\mathbf{N^{corr}_{cand}} \stackrel{fil}{=} \{(\mathbf{n} \in \mathbf{N_{cand}}) | \hat{\mathbf{C}}_\mathbf{p}(\mathbf{n}) > \theta\} \quad (2)$$

Because norms are subjective in nature and language models have the risk of hallucination in their output predictions, we further safeguard norm discovery with a $\textbf{P}\text{ROMPT}_\mathbf{grd}(.)$ operator for determining whether the hypothesized norm discovery can be groundable to its situation premise. We draw inspirations from the explainable NLI setting (Camburu et al., 2018), and formulate grounding by the following natural language template: *"Explain whether the conversation <**entails**, **contradicts**, or **is irrelevant with**> the given norm"*. The output from $\textbf{P}\text{ROMPT}_\mathbf{grd}(.)$ consists of the grounding verdict, $\mathbf{G_v}$, along with the explanation, $\mathbf{G_{expl}}$ (see Figure 1b for example). We further derive a confidence score for the grounding relevance:

$$\hat{\mathbf{G}}_\mathbf{p} = \frac{\mathbf{P}_{\max}(\mathbf{G_v} \neq 0)}{\mathbf{P}_{\max}(\mathbf{G_v} \neq 0) + \mathbf{P}(\mathbf{G_v} = 0)} \quad (3)$$

and filter out the norm candidates with grounding score below a tunable threshold, $\gamma = 0.6$.

$$\mathbf{N^{fil}_{cand}} \stackrel{fil}{=} \{(\mathbf{n} \in \mathbf{N^{corr}_{cand}}) | \hat{\mathbf{G}}_\mathbf{p}(\mathbf{n}) \geq \gamma\} \quad (4)$$

Finally, when populating **NORMSKB** with new norm discoveries, we perform self-verification only for the norms that are not duplicates of existing norms in the norms library. We flag norms as duplication when their BERT(n) embeddings (Devlin et al., 2019) exceed a threshold of cosine similarity with any previously discovered norm. The threshold is empirically set to $\sigma = 0.95$.

$$\mathbf{N} = \mathbf{N} \cup \{(\mathbf{n} \in \mathbf{N}, \mathbf{n}' \in \mathbf{N^{fil}_{cand}}) \;|\; cos(BERT(\mathbf{n}), BERT(\mathbf{n}')) < \sigma\} \quad (5)$$

4 Evaluation and Results

4.1 Intrinsic Norm Discovery Evaluation

Baselines We include the following relevant baseline methods for norm discovery:

- **NMT**$_{gen}$: This is a GPT2-XL trained on SOCIALCHEM101 (Forbes et al., 2020).
- **SOCIALCHEM**$_{rtrv}$: This retrieves the most relevant SOCIALCHEM101 rule-of-thumbs for a given dialogue, based on their embeddings encoded from pre-trained BERT (Devlin et al., 2019).
- **pMT**$_{gen}$: This is a generator trained on the MORAL INTEGRITY CORPUS (MIC) (Ziems et al., 2022).

	Relevance	Well-Formedness	Correctness	Insightfulness	Relatableness
SocialChem$_{rtv}$	3.8	4.0	**3.9**	3.8	3.9
NMT$_{gen}$	3.4	3.9	3.6	3.4	3.7
MIC$_{rtv}$	2.2	3.1	3.3	3.4	2.5
pMT$_{gen}$	2.2	3.0	3.0	3.0	2.5
T0$_{pp}$	2.7	2.0	2.0	2.1	2.1
NormSage$_{base}$	3.0	2.8	2.83	2.8	3.6
NormSage$_{frame}$	**<u>4.5</u>**	**<u>4.5</u>**	**<u>4.6</u>**	**<u>4.2</u>**	**<u>4.7</u>**
NormSage$_{indc}$	**3.8**	**4.5**	3.8	**3.9**	**3.9**

Table 3: Likert scale (1-5) results, averaged over 100 data samples.

- **MIC**$_{rtrv}$: This retrieves the most relevant MIC rules-of-thumb for a given dialogue, based on their embeddings encoded from pre-trained BERT.

- **T0**$_{pp}$: This is a T5 model trained on tasks formulated as natural language prompts (Sanh et al., 2022). It is 16x smaller than GPT-3.

For our proposed framework, we include the **NormSage**$_{base}$, which contains no structural enhancement in the prompt, as well as **NormSage**$_{frame}$ and **NormSage**$_{indc}$, which incorporates frame guidelines and structured indicators, respectively.

Metrics We measure norm discovery from in-situ conversations on a Likert scale of 1-5, with 1 as "awful" and 5 as "excellent", through the following evaluation criteria:

- ***Relevance***: can we see that the norm is inspired from the situation (lower bound on norm applicability).

- ***Well-Formedness***: how well is the norm structured – is the norm self-contained, and does it include *both* a judgment of acceptability or occurrence, *and* an action or societal/cultural phenomena that is assessed.

- ***Correctness***: to the best of their knowledge, would people agree that the described norm holds true?

- ***Insightfulness***: does the norm convey englightening understanding about what's considered acceptable and standard in the society that pertain to the conversation scenario.

- ***Relatableness***: how well does the norm balance vagueness against specificity, so that it can generalize across multiple situations (e.g., "It is rude to be selfish.") without being too specific (e.g., "It is rude not to share your mac'n'cheese with your younger brother.")

Setting We crowdsource Amazon Mechanical Turk for human assessment. Each HIT ("submit" task) consists of a dialogue scenario, one of the metrics to assess, and three sets of norms, each representing the norms discovered from a generation or retrieval approach, de-identified. To further constrain the amount of norms within each set of discovery method included for manual assessment, we select only the first three norms generated or retrieved for a given dialogue. Following crowdsourcing guidelines outlined in (Sheehan, 2018), we provide definitions and detailed examples for each assessment metric. Workers undergo a vetting process before working on norm evaluation, including a qualification criteria of $\geq$95% HIT rate, and checks that they understand what norms are and what the given assessment metrics are about. We assign ten workers per example, and reject poor quality hits, such as the hits from workers who leave all Likert scale entries to the default value or rate specific event instances (non-norms) with a high score. Workers take 1-2 minutes per norm comparison task, and HITs are rewarded \$0.34 each. For norm rating, the inter-annotator agreement had a Cohen's kappa of 0.41, which is moderate agreement, but this is expected since annotators may calibrate their scores differently on the Likert scale (Ziems et al., 2022).

Results We show our norm discovery intrinsic evaluation results in Table 3. We can see that our proposed norm discovery approach, **NormSage**, outperforms baselines across all dimensions when enhanced with either frame or structured indicators. A major limitation of baseline approaches

Culture-Specific Norm Discoveries	**Source**
In **Pakistani** culture, it is common for women to wear headscarves. In **Pakistani** culture, it is not uncommon for the bride and groom to not meet each other until the wedding day.$^{\bullet}$ In **Pakistani** culture, it is more common for marriages to take place within the same religion.	Citizen Khan (C.K.)
In **India**, it is considered polite to always offer food and drink to guests, even if they decline.$^{\diamond}$ In **India**, people often eat with their hands instead of with utensils.	Outsourced (O)
In **Taiwanese** culture, it is more common to have a heavier lunch, such as rice and vegetables.$^{\circ}$ In **Taiwanese** culture, it is common for people to take their shoes off when entering a home.	Fresh off the Boat (F.B.)
In the **African-American** culture, it is common for people to listen to music with a strong beat. In **African-American** culture, it is proper to show respect for your elders by calling them by their title (Mr., Mrs., Miss, Ms., etc.)	Blackish (B)
In **British** culture, it is considered normal for the bride and groom to meet each other before the wedding day.$^{\bullet}$ In **American** culture, it is common to have a light lunch, such as a salad or sandwich.$^{\circ}$ In **America**, it is more common to just let guests decline if they don't want anything.$^{\diamond}$	C.K., O, F.B.

Table 4: Visualization of culture-specific norm discovery examples. We denote the pairs of contrasting norms across cultures with special symbols ($\diamond, \circ, \bullet$).

is poor portability to conversation domains. The performance of SOCIALCHEM$_{rtv}$ and **MIC**$_{rtv}$ shows that simply retrieving pre-annotated norms results in the norms being less relevant and insightful for new conversations. Compared to the retrieval baselines, the generation baselines, **NMT**$_{gen}$ and **PMT**$_{gen}$, perform even worse. This suggests that the domain gap in situation context between curated Reddit post headers (previous works) and in-situ conversations (current task) poses an even greater bottleneck for norm discovery here. NORMSAGE overcomes the challenges in domain portability through operationalizing zero-shot language model prompting for conversation reasoning and norm discovery.

4.1.1 Quality of Multi-Lingual Norm Discovery:

The norms discovered from Chinese conversations are high-quality in detail and correctness, as visualized in Fig 3. We also perform a investigation on the quality of norms discovered from Chinese conversations compared to norms discovered from English conversations. We measure the standalone correctness and insightfulness of norms discovered from multi-lingual setting, on a 1-5 Likert scale. The results in Table 5 indicate that norms discovered from English data are rated slightly higher in insightfulness ($\leq$1%) but lower in correctness ($\leq$3%), potentially due to the dialogue nature. The English sources involve movies and show, which tend to be more creative (insightful) and less formal (biased dialogues may lower norm correctness).

Figure 3: Example norms discovered from Chinese conversations.

	EN Conv. Norm	CN Conv. Norm
Insightfulness	4.54	4.47
Correctness	4.65	4.79

Table 5: A Likert-scale comparison of norms discovered from English (EN) vs Chinese (CN) conversation.

4.1.2 Quality of Culture-Specific Norm Discovery:

To the best of our knowledge, we are the first work in discovering culture-specific norms. To evaluate the correctness of culture-specific norms, we design a pairwise culture comparison setting.

Specifically, we run a pretrained **BART-LARGE-MNLI** model (Lewis et al., 2020; Williams et al., 2018) on pairs of norm from different cultures, and randomly select 10 pairs of norm that are determined as "contradiction" with each other, for each of the cross-culture scenarios in our dataset. Then, we mask the culture identities in the pairs of norm, and ask human annotators familiar with both cultures to identity which culture each norm belongs to from binary options. The results, as shown in Table 6, indicate that the culture-specific norms discovered from NORMSAGE are promising, with human annotators achieving 80% identification accuracy. Some of the error cases in culture comparison of norm discoveries may be due to subjectivity of the assessment task (for example, whether the White or Black cultural group in America is more likely to discuss controversial topic).

Culture Comparison	ID (%)	Rating (1-5)
American vs E. Asian	74	4.2
American vs Indian	82	4.0
Western vs Muslim	91	4.5
White vs Black	73	4.0
Average	80	4.1

Table 6: Culture-specific norm evaluation via cross-culture identification (ID) comparison.

4.2 Extrinsic Norm Grounding and Explanation Evaluation

Norm grounding is utilized in the self-verification process of norm discovery by **NORMSAGE**. This subtask naturally extends to the online instantiation of norm adherence and violation in downstream applications. Thus, we perform extrinsic evaluation on model performance in the explainable grounding of norms on dialogue situations.

4.2.1 Baselines

We compare **NORMSAGE** with the following:

- **BART-MNLI**: This is BART-LARGE (Lewis et al., 2020) pretrained on the Multi-genre NLI corpus (Williams et al., 2018).
- **BART-DIALNLI**: This is the BART-LARGE model pretrained on the Dialogue NLI corpus (Welleck et al., 2018).
- **T5-eSNLI**: This is T5 (Raffel et al., 2020) trained on the explainable NLI, or e-SNLI, dataset (Camburu et al., 2018).
- **T0$_{pp}$**: This has been introduced in Sec 4.1.
- **Human-Labeling**: We also include human-written grounding explanations, as an upper-bound for comparison.

4.2.2 Metrics

We measure norm grounding in terms of three-class classification accuracy (Acc) and area under the ROC-curve (AUC). The latter takes into account the confidence score, or probability, of grounding. In addition, we evaluate the natural language explanations of grounding predictions using human assessment, on a 1-5 Likert scale.

4.2.3 Setting

We perform a class-balanced random sampling of 100 ($\mathbf{X}_{dial_i}$, $norm_{cand}$) pairs. The class is determined prelimarily from NORMSAGE as described in Sec 3.2. Next, we obtain crowdsourced annotations for the gold standard grounding verdict and explanations in this subset of the data.

4.2.4 Results

The grounding results, as summarized in Table 7, show that **NORMSAGE** outperforms all baselines in grounding verdict inference. In addition, grounding explanations from our **NORMSAGE** framework is preferred over human-written ones in approximately 40% of the cases, which suggests that our automatic explanation generation approach is strong and competitive. We provide a visualization of the norm discoveries, grounding, and explanation in Table 8.

	Acc (%)	AUC	Expl (1-5)
BART-MNLI	46	34.0	N/A
BART-DIALNLI	42	37.3	N/A
T5-eSNLI	47	51.2	3.33
T0	29	34.1	3.26
NORMSAGE	**81**	**95.4**	3.49
Human-Labeling	N/A	N/A	**3.57**

Table 7: Instantiation accuracy and likert scale results.

4.3 Resource Contribution

We discovered over 20,500 unique norms, of which 1,250 are culture-specific. On average,

Dialogue Situation	Discovered Norms	Grounding Explanation	Ground. Verdict
Dave: No. Definitely booked. **Mr. Khan**: What?! Do know who I am? Hello! Mr Khan, community leader! Next President of Sparkhill Pakistani Business Association! **Dave**: I'm sorry **Mr. Khan**: Right, that's it. I want to speak to the proper manager. **Dave**: I am the property manager.	It's important to listen to others and give them a chance to speak.	Mr. Khan is not listening to Dave and he is not giving Dave a chance to speak.	-1
Beckett: Sure I can, until a jury tells me otherwise. **Creason**: You are wasting my time. Detective, look, I told you exactly what I was doing last night. **Beckett**: Right. You were at the club. They said that you made quite the entrance. ...	It is generally considered impolite to make lewd comments.	What's spoken by Creason is irrelevant with the norm.	0
... **Jessica**: Well, those kids, they just don't know, that's all. It just – it just take time to get used to something different. **Eddie**: I hate it here! I want to go back to D.C. **Jessica**: Eddie, that's not possible. We are here now. We have to make the best of it. Like I am doing with this neighbor woman. You think I like pretending Samantha isn't carrying a baggie of dog poops in her hand? No! I don't like this! We all see the poops there! It's rolling around But I am trying! You have to try, too.	It is also considered polite to try to make the best of a situation, even if you do not like it	The mother is trying to make the best of the situation even though she does not like it	1

Table 8: Norm grounding example results, randomly sampled for each class from *{Contradict (-1), Irrelevant (0), Entail (1)}*. We underline the utterance-level provenance of the grounding instance, in cases which entailment or contradiction are found.

NORMSAGE discovers norms at a rate of 8.6 seconds per dialogue, and performs norm grounding at a rate of 3.8 seconds per dialogue. This is over 10x faster the human annotation efforts.

5 Related Work

The domain of norms is closely related to behavioral psychology and moral judgement. Early studies investigated the pragmatic cooperative principles (Grice, 1975), politeness implicatures (Kallia, 2004), and relationship between norms and law (Posner, 2009) governing human behavior. As judgements of behavior are communicated through linguistics, (Graham et al., 2009) introduced a lexicon of evocative words based on moral foundation theory, which later attempts utilize for predicting the moral value from text messages (Lin et al., 2018; Mooijman et al., 2018). Recent approaches explore modeling moral and ethical judgement of real-life anecdotes from Reddit (Emelin et al., 2021; Sap et al., 2019a; Lourie et al., 2021; Botzer et al., 2022), with DELPHI (Jiang et al., 2021a) unifying the moral judgement prediction on these related benchmarks. Related is another line of work modeling legal judgement on judicial corpora (Chalkidis et al., 2022).

Norm discovery is a unique, emerging task, which aims to catalogue the underlying principles behind behavioral judgements, and can be seen as similar to distilling reactions, explanations, and implications from situations (Vu et al., 2014; Ding and Riloff, 2016; Rashkin et al., 2018; Sap et al., 2019b). Forbes et al. (2020); Ziems et al. (2022) are the main existing norm discovery approaches. Each presents a large-scale catalogue of manually curated rule-of-thumbs from Reddit post headers, and trains a language model to generate rule-of-thumbs based on this data. In contrast, our work focuses on norm discovery from conversations on-the-fly and without needing manual curation.

Modeling the social and moral dynamics in human interaction and communication have diverse applications, such as the detection of cyberbullying (Van Hee et al., 2015), bipartisan news framing (Fulgoni et al., 2016), social media post fact *vs.* fiction (Volkova et al., 2017), emotions (Zadeh et al., 2018; Yu et al., 2020), and situational QA (Gu et al., 2022). In particular, discovering norms is essential for *explicitly* detecting norm adherence and violations instances (our work), as well as *implicitly* guiding dialogues (Ziems et al., 2022).

From a technical perspective, our norm discovery approach based on language model prompting and knowledge elicitation can be seen as a form of prompt engineering (Le Scao and Rush, 2021), where we prefix a question with an elaborated scene. The norm grounding with explanation task is intuitively similar to the explainable natural language inference problem setting (Welleck et al., 2018; Wiegreffe et al., 2021). Our proposed framework, NORMSAGE, achieves norm discovery and grounding without intensive prompt-tuning (Jiang et al., 2021b) or finetuning (Forbes et al., 2020; Ziems et al., 2022).

6 Conclusions and Future Work

We present **NORMSAGE**, a framework for conversation grounded norm discovery through language model prompting and self-verification. It achieves greater depth and breadth in detailing the underlying rules of acceptable behavior and expectations for a wide-range of dialogue situations across social and cultural groups, compared to baselines based on static knowledge, such as the Reddit forum data curated from crowd-sourced annotators. It is also capable of discovering high-quality norms from multi-lingual conversation, and norms with culture-specific awareness, which are novel task settings. Finally, we are the first work to achieve automated natural language grounding explanations for interpretable norm discovery and norm violation detection, comparable to human-written grounding explanations. For future research directions, we believe it is meaningful to explore direct norm discovery from cross-modal settings, leveraging audio-visual cues, and performing finer-grained norm categorization.

7 Ethical Considerations

In this work, our norm discovery process makes use of GPT3 as a strong pre-trained language model to elicit groundable knowledge about the rules and judgements of acceptable behavior from human dialogue interactions. We recognize that social, socio-cultural, and moral norms may shift with context over time. Our discovery of norms applies to the time period that aligns with the conversation scenario in which a norm is discovered from. We further point out that the GPT3 model acquired its implicit knowledge from ultra large-scale data, and has added in mechanisms to address bias (Solaiman and Dennison, 2021). Nevertheless, all computational models still come with a risk of potential bias. We encourage researchers and practitioners to exercise caution and check-guards in their endeavors.

We recognize that the automatic generation of norms and judgements, could be seen as normative and authoritative (Talat et al., 2021; Ziems et al., 2022). We emphasize that we do not treat the discovered norms as global or universally binding. The norms are not designed to form a cohesive and universal ethical system, but rather to provide a set of discrete intuitions and principles to help differentially explain the underlying assumptions that exist latently. The present work supports an explainable system to verify whether a discovered norm can be sufficiently grounded to its data source, and the relation characteristic (entail vs. contradict). Moderation efforts can appear at a later stage, handled by domain experts who may interface with our transparent and flexible system.

Risks and Mitigations

Our task involves source data that may contain explicit conversations about race, gender, religion, etc. We recognize the emotional burden that this presents to annotators (Roberts, 2016). In mitigation, we include the following content warning in the header of each task: *This HIT may contain text that disturbs some workers. If at any point you do not feel comfortable, please feel free to skip the HIT or take a break.* The study has been thoroughly reviewed and approved by a national level internal review board.

The resources and findings presented in this work are intended for research purposes only. To ensure proper, rather than malicious, application of dual-use technology, we require users of our norm discovery data to complete a Data Usage Agreement that we link in our project repository. We also intend to make our software available as open source for public auditing, and explore measures to protect vulnerable groups.

References

Zsuzsanna Abrams, Nancy Guilloteau, and Lia Plakans. 2022. Sociocultural norms | foreign language teaching methods. Accessed: 2022-06-01.

Nicholas Botzer, Shawn Gu, and Tim Weninger. 2022. Analysis of moral judgment on reddit. *IEEE Transactions on Computational Social Systems*.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020a. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020b. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Oana-Maria Camburu, Tim Rocktäschel, Thomas Lukasiewicz, and Phil Blunsom. 2018. e-snli: Natural language inference with natural language explanations. In *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc.

Ilias Chalkidis, Abhik Jana, Dirk Hartung, Michael Bommarito, Ion Androutsopoulos, Daniel Katz, and Nikolaos Aletras. 2022. LexGLUE: A benchmark dataset for legal language understanding in English. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4310–4330, Dublin, Ireland. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Haibo Ding and Ellen Riloff. 2016. Acquiring knowledge of affective events from blogs using label propagation. In *Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence*, AAAI'16, page 2935–2942. AAAI Press.

Denis Emelin, Ronan Le Bras, Jena D. Hwang, Maxwell Forbes, and Yejin Choi. 2021. Moral stories: Situated reasoning about norms, intents, actions, and their consequences. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 698–718, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Maxwell Forbes, Jena D. Hwang, Vered Shwartz, Maarten Sap, and Yejin Choi. 2020. Social chemistry 101: Learning to reason about social and moral norms. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 653–670, Online. Association for Computational Linguistics.

Dean Fulgoni, Jordan Carpenter, Lyle Ungar, and Daniel Preoţiuc-Pietro. 2016. An empirical exploration of moral foundations theory in partisan news sources. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 3730–3736, Portorož, Slovenia. European Language Resources Association (ELRA).

Jesse Graham, Jonathan Haidt, and Brian Nosek. 2009. Liberals and conservatives rely on different sets of moral foundations. In *Journal of Personality and Social Psychology*, pages 1029–1046. Brill.

Herbert P Grice. 1975. Logic and conversation. In *Speech acts*, pages 41–58. Brill.

Yuling Gu, Bhavana Dalvi, and Peter Clark. 2022. DREAM: Improving situational QA by first elaborating the situation. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1115–1127, Seattle, United States. Association for Computational Linguistics.

Liwei Jiang, Jena D Hwang, Chandra Bhagavatula, Ronan Le Bras, Maxwell Forbes, Jon Borchardt, Jenny Liang, Oren Etzioni, Maarten Sap, and Yejin Choi. 2021a. Delphi: Towards machine ethics and norms. *arXiv preprint arXiv:2110.07574*.

Zhengbao Jiang, Jun Araki, Haibo Ding, and Graham Neubig. 2021b. How can we know when language models know? on the calibration of language models for question answering. *Transactions of the Association for Computational Linguistics*, 9:962–977.

Alexandra Kallia. 2004. Linguistic politeness: The implicature approach. *Multilingua*, 23.

Teven Le Scao and Alexander Rush. 2021. How many data points is a prompt worth? In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2627–2636, Online. Association for Computational Linguistics.

Jie Lei, Licheng Yu, Mohit Bansal, and Tamara Berg. 2018. TVQA: Localized, compositional video question answering. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 1369–1379, Brussels, Belgium. Association for Computational Linguistics.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7871–7880, Online. Association for Computational Linguistics.

Ying Lin, Joe Hoover, Gwenyth Portillo-Wightman, Christina Park, Morteza Dehghani, and Heng Ji. 2018. Acquiring background knowledge to improve moral value prediction. In *Proc. The 2018 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM2018)*.

Pengfei Liu, Weizhe Yuan, Jinlan Fu, Zhengbao Jiang, Hiroaki Hayashi, and Graham Neubig. 2021. Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing. *arXiv preprint arXiv:2107.13586*.

Nicholas Lourie, Ronan Le Bras, and Yejin Choi. 2021. Scruples: A corpus of community ethical judgments on 32,000 real-life anecdotes. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35(15), pages 13470–13479.

Marlon Mooijman, Joe Hoover, Ying Lin, Heng Ji, and Morteza Dehghani. 2018. Moralization in social networks and the emergence of violent protests. *Nature Human Behavior [**June 2018 Cover**]*.

Fabio Petroni, Tim Rocktäschel, Sebastian Riedel, Patrick Lewis, Anton Bakhtin, Yuxiang Wu, and Alexander Miller. 2019. Language models as knowledge bases? In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 2463–2473, Hong Kong, China. Association for Computational Linguistics.

Eric Posner. 2009. *Law and social norms*, chapter 1-4. Havard University Press.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, Peter J Liu, et al. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *J. Mach. Learn. Res.*, 21(140):1–67.

Hannah Rashkin, Maarten Sap, Emily Allaway, Noah A. Smith, and Yejin Choi. 2018. Event2Mind: Commonsense inference on events, intents, and reactions. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 463–473, Melbourne, Australia. Association for Computational Linguistics.

Sarah T Roberts. 2016. Commercial content moderation: Digital laborers' dirty work. In *The Intersectional Internet: Race, Sex, Class and Culture Online*. Peter Lang Publishing.

Victor Sanh, Albert Webson, Colin Raffel, Stephen Bach, Lintang Sutawika, Zaid Alyafeai, Antoine Chaffin, Arnaud Stiegler, Arun Raja, Manan Dey, M Saiful Bari, Canwen Xu, Urmish Thakker, Shanya Sharma Sharma, Eliza Szczechla, Taewoon Kim, Gunjan Chhablani, Nihal Nayak, Debajyoti Datta, Jonathan Chang, Mike Tian-Jian Jiang, Han Wang, Matteo Manica, Sheng Shen, Zheng Xin Yong, Harshit Pandey, Rachel Bawden, Thomas Wang, Trishala Neeraj, Jos Rozen, Abheesht Sharma, Andrea Santilli, Thibault Fevry, Jason Alan Fries, Ryan Teehan, Teven Le Scao, Stella Biderman, Leo Gao, Thomas Wolf, and Alexander M Rush. 2022. Multitask prompted training enables zero-shot task generalization. In *International Conference on Learning Representations*.

Maarten Sap, Saadia Gabriel, Lianhui Qin, Dan Jurafsky, Noah A Smith, and Yejin Choi. 2019a. Social bias frames: Reasoning about social and power implications of language. *arXiv preprint arXiv:1911.03891*.

Maarten Sap, Ronan LeBras, Emily Allaway, Chandra Bhagavatula, Nicholas Lourie, Hannah Rashkin, Brendan Roof, Noah A Smith, and Yejin Choi. 2019b. Atomic: An atlas of machine commonsense for if-then reasoning. In *AAAI*.

Kim Bartel Sheehan. 2018. Crowdsourcing research: data collection with amazon's mechanical turk. *Communication Monographs*, 85(1):140–156.

Irene Solaiman and Christy Dennison. 2021. Process for adapting language models to society (palms) with values-targeted datasets. *Advances in Neural Information Processing Systems*, 34:5861–5873.

Zeerak Talat, Hagen Blix, Josef Valvoda, Maya Indira Ganesh, Ryan Cotterell, and Adina Williams. 2021. A word on machine ethics: A response to jiang et al.(2021). *arXiv preprint arXiv:2111.04158*.

Cynthia Van Hee, Els Lefever, Ben Verhoeven, Julie Mennes, Bart Desmet, Guy De Pauw, Walter Daelemans, and Veronique Hoste. 2015. Detection and fine-grained classification of cyberbullying events. In *Proceedings of the International Conference Recent Advances in Natural Language Processing*, pages 672–680, Hissar, Bulgaria. INCOMA Ltd. Shoumen, BULGARIA.

Svitlana Volkova, Kyle Shaffer, Jin Yea Jang, and Nathan Hodas. 2017. Separating facts from fiction: Linguistic models to classify suspicious and trusted news posts on Twitter. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 647–653, Vancouver, Canada. Association for Computational Linguistics.

Hoa Trong Vu, Graham Neubig, Sakriani Sakti, Tomoki Toda, and Satoshi Nakamura. 2014. Acquiring a dictionary of emotion-provoking events. In *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics, volume 2: Short Papers*, pages 128–132, Gothenburg, Sweden. Association for Computational Linguistics.

Sean Welleck, Jason Weston, Arthur Szlam, and Kyunghyun Cho. 2018. Dialogue natural language inference. *arXiv preprint arXiv:1811.00671*.

Sarah Wiegreffe, Jack Hessel, Swabha Swayamdipta, Mark Riedl, and Yejin Choi. 2021. Reframing human-ai collaboration for generating free-text explanations. *arXiv preprint arXiv:2112.08674*.

Adina Williams, Nikita Nangia, and Samuel Bowman. 2018. A broad-coverage challenge corpus for sentence understanding through inference. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1112–1122. Association for Computational Linguistics.

Wenmeng Yu, Hua Xu, Fanyang Meng, Yilin Zhu, Yixiao Ma, Jiele Wu, Jiyun Zou, and Kaicheng

Yang. 2020. CH-SIMS: A Chinese multimodal sentiment analysis dataset with fine-grained annotation of modality. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 3718–3727, Online. Association for Computational Linguistics.

AmirAli Bagher Zadeh, Paul Pu Liang, Soujanya Poria, Erik Cambria, and Louis-Philippe Morency. 2018. Multimodal language analysis in the wild: Cmu-mosei dataset and interpretable dynamic fusion graph. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2236–2246.

Caleb Ziems, Jane Yu, Yi-Chia Wang, Alon Halevy, and Diyi Yang. 2022. The moral integrity corpus: A benchmark for ethical dialogue systems. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3755–3773, Dublin, Ireland. Association for Computational Linguistics.